Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface.

Article (Published Version)

Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface

A Rosemary Tate,[1] Natalia Beloff,[1] Balques Al-Radwan,[1] Joss Wickson,[2] Shivani Puri,[3] Timothy Williams,[3] Tjeerd Van Staa,[3,4,5] Adrian Bleach[2]

[1]Department of Informatics, University of Sussex, Brighton, UK
[2]Dataline Software Ltd, Brighton, UK
[3]CPRD, Medicines and Healthcare Products Regulatory Agency, London, UK
[4]London School of Hygiene & Tropical Medicine, London, UK
[5]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

Correspondence to
Dr A Rosemary Tate, Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK; rosemary@sussex.ac.uk

Received 9 April 2013
Revised 15 October 2013
Accepted 5 November 2013

ABSTRACT

Objective UK primary care databases, which contain diagnostic, demographic and prescribing information for millions of patients geographically representative of the UK, represent a significant resource for health services and clinical research. They can be used to identify patients with a specified disease or condition (phenotyping) and to investigate patterns of diagnosis and symptoms. Currently, extracting such information manually is time-consuming and requires considerable expertise. In order to exploit more fully the potential of these large and complex databases, our interdisciplinary team developed generic methods allowing access to different types of user.

Materials and methods Using the Clinical Practice Research Datalink database, we have developed an online user-focused system (TrialViz), which enables users interactively to select suitable medical general practices based on two criteria: suitability of the patient base for the intended study (phenotyping) and measures of data quality.

Results An end-to-end system, underpinned by an innovative search algorithm, allows the user to extract information in near real-time via an intuitive query interface and to explore this information using interactive visualization tools. A usability evaluation of this system produced positive results.

Discussion We present the challenges and results in the development of TrialViz and our plans for its extension for wider applications of clinical research.

Conclusions Our fast search algorithms and simple query algorithms represent a significant advance for users of clinical research databases.

BACKGROUND AND SIGNIFICANCE

Opportunities for the secondary use of routinely collected data for research purposes have increased enormously in recent years due to the wide uptake and advances in technology; for example, electronic health records (EHR). The continued expansion of large databases of patient records, often linked to other sources of data, has greatly enhanced the possibilities for using these records for health services and clinical research.

Most people in the UK are registered with a general practitioner (GP) who is the first port of call if they have a health problem. GPs act as the 'gatekeeper' to the National Health Service (NHS), which is free at the point of use. They deal with minor ailments locally, but refer patients for further tests or care if the problem appears to be more serious or cannot be treated within the practice.

All general practice encounters are recorded electronically and practitioners are encouraged to make these records available for research. The Clinical Practice Research Datalink (CPRD) database represents the largest collection of anonymized primary care patient records in the world.[1] Data are collected from 5.5 million currently registered patients (approximately 9% of the UK population). These data are used worldwide by academics, governments and the pharmaceutical industry for health services and clinical research.

However, extracting relevant information (feature extraction or phenotyping[2]) from this database is often difficult and time-consuming. Each patient may have thousands of records often with multiple events occurring on the same day. Although these records are rich in information, in common with most data collected for non-research purposes, records are variable in quality and may be missing or incomplete. The present use of EHR databases requires programing expertise and understanding of the nuances of the coding systems. Queries may take hours or even days to run, and once obtained the only way most researchers can view the results is to scroll through hundreds of records in tabular form.

OBJECTIVES

Our goal is to develop methods that will make the information in these databases accessible to different types of users, including those with little expertise or those with little understanding of the underlying data models and/or nuanced coding schemes. In this project we developed methods to allow users to identify patients suitable for further screening for recruitment into randomized controlled trials (RCT) within general practices. TrialViz is a simple online intuitive interface to the large and complex data held within CPRD. With simple yet powerful interactive tools to build complex trial protocols, TrialViz enables users to select suitable GP practices based on two criteria: suitability of the patient base for the intended study (phenotyping) and practice-based measures of the quality of data recording. Demographic and clinical parameters (such as test results) supplement the coded data to provide feasibility data in terms of practice quality and potential recruitment rates. We aim to represent the disease prevalence components in real time, thus enabling a flexible, rapid analysis of study feasibility under a range of different assumptions and choices. In this paper we describe the challenges and results in the development of

To cite: Tate AR, Beloff N, Al-Radwan B, *et al*. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2013-001847

the system, our usability evaluation of the TrialViz tool and our plans for its further development for wider applications of clinical research.

MATERIALS AND METHODS

The database

CPRD (formerly the General Practice Research Database) is the world's largest validated computerized database of anonymized longitudinal medical records for primary care.[1] Data comprise approximately 14 million patients with around 5.4 million of these being currently alive and registered from 660 primary care practices spread throughout the UK. CPRD is used worldwide for research by the pharmaceutical industry, clinical research organizations, regulators, government organizations and leading academic institutions.

Records are derived from a widely used GP software system (VISION) and contain complete prescribing and coded diagnostic and clinical information as well as information on tests requested, laboratory results and referrals made at or following on from each consultation (figure 1). Each clinical event has a code assigned by the GP using the Read coding system, V.2.[3] Read codes are a hierarchical recording system used to record clinical summary information and are not limited to diagnostic and procedural codes, but also include codes for symptoms, test results, screening, history and other areas. CPRD currently includes approximately 4 billion records originally extracted from primary care practices, in 1/2 terabyte of relational data. This data volume will increase over time with the introduction of further sources, including cancer registries (with information on the grade and stage of the cancer), death and birth registries, hospital episode statistics and socioeconomic class information.

Methods

Our multidisciplinary team of data analysts, epidemiologists, statisticians, graphical designers, software engineers, and computer scientists developed:

1. Data abstraction tools – presented via an intuitive user interface whereby users can upload or select codes/codelists and create rules or queries for selecting the patients of interest.
2. Data extraction tools for running queries using the data abstraction rules in near real time.
3. A protocol for measuring data quality for each practice and its fitness for a particular study.
4. Visualization tools to investigate the results.

These tools have been developed as part of a web-based system 'TrialViz', funded by the UK Technology Strategy Board. This system will enable research organizations and pharmaceutical companies to undertake feasibility assessments to identify suitable patients within GP practices for recruitment into clinical trials, based on the available number of patients and the data quality of the practice.

Data abstraction

One of the main challenges of working with UK GP data is the number and complexity of the queries that need to be carried out in order to sift out relevant information from the mass of (mainly irrelevant) data. There may be thousands of records for each patient, and there may be numerous codes for each disease or type of symptom representing essentially the same thing. For example, there are over 200 codes for diabetes and 40 codes representing 'abdominal pain'. Before carrying out an analysis using GP data the user must draw up code lists, that is, a comprehensive set of condition-specific medical or drug codes that can be used to search patient medical/clinical and prescription records.[4] After constructing the code list, the user must develop a set of rules for extracting patients and events of interest and write queries to merge all the relevant tables and extract records for particular patients or events. This non-trivial task requires considerable expertise. Many analysts use statistical software to manage and extract the information, a minority use a database query language such as SQL. Because running these queries may take hours, or even days, analysts normally work on a small subset of the data.

Data extraction

Another major challenge is the process time involved in running queries. For data to be explored in an interactive and robust manner, results of queries should be returned in real or near real time. Most SQL-based systems that exist at present are not capable in general of doing this. SQL is known to be erratic when working with large datasets, that is, if the parser decides to execute a query in an inefficient manner, large data volumes tend to exacerbate the delinquent behavior, resulting in query execution time of several hours or days rather than expected seconds. Thus it is often not possible to estimate in advance how long each query will take. It is also important to note that the ability to conduct large dataset research in real time in principle is a very recent phenomenon, because storage capabilities

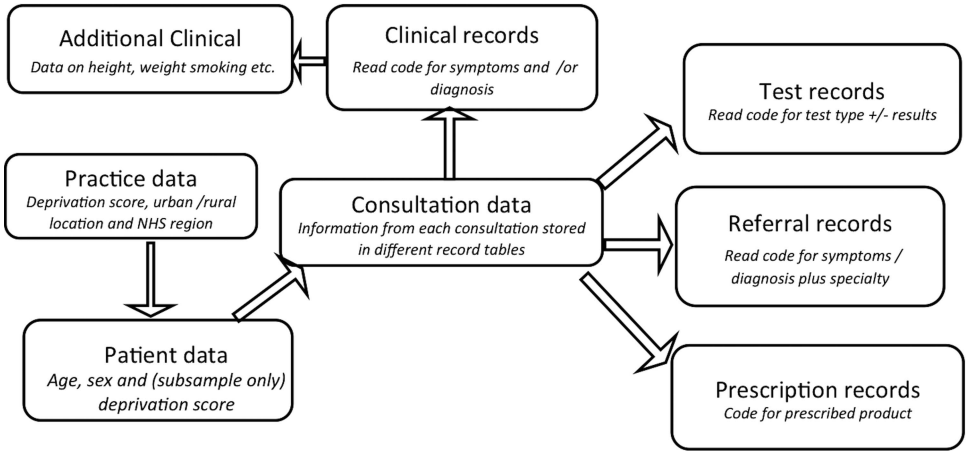

Figure 1 Primary care database overview.

necessary to do so, such as solid state drives, only recently became affordable for general research use.

Data quality assurance

In any application based on data, it is very important to ensure that the data are of high quality. For example, it is unethical to enrol patients into an experiment in which poor quality data collection could lead to invalid results. Poor quality data lead to poor searches with low efficiency for screening visits. Thus, data quality, which we define as fitness for use, is a 'sine qua non' for the use of EHR in RCT. Investigation and monitoring of data quality is a key component of the TrialViz project. Although many books and articles have been written about data quality in general, there exists no commonly accepted methodology for defining and comparing data quality in medical databases used for research.[5 6] Thus developing such a protocol formed an important part of this work.

The work was carried out in several iterations, starting with a preliminary investigation of basic and easily measurable criteria. After a review of the literature, we selected a suitable framework for monitoring data quality (that of the UK audit commission) and used this to help define more complex and study-specific aspects of data quality. The results of this investigation were used to develop a protocol for characterizing and monitoring the data quality of practices contributing data to the CPRD (manuscript in preparation). Protocol development was informed by the results of an investigation of data quality in all 629 practices contributing data to the CPRD each year between 2000 and 2010. We extracted simple ratios (ie, number of desired outcomes/total number of outcomes)[7] on general measures such as missing dates or codes, and study-specific measures related to diabetes and coronary heart disease, for example, completeness of relevant test results. The distributions of these variables (the ratios) and their interrelationships were investigated using summary statistics, graphs (histograms, boxplots and scatterplots) and correlation analysis.

User evaluation

An inherent part of any software application development is a strong user involvement at all stages of the development cycle, from the conception of requirements specification, through the prototype testing to the usability evaluation. The TrialViz team has ensured a continuous stakeholder involvement throughout the development process, with regular user group meetings, at which academic medical researchers, medical statisticians, primary care research network members, clinical trial facilitators and leading pharmaceutical companies were regularly steering the development towards greater usability of the tool. In addition to the scheduled usability testing for the TrialViz primary purpose (feasibility studies for randomized clinical trials), the tool was also subjected to a usability study on a group of potential users from a variety of backgrounds: researchers (five in each group) in epidemiology, medical statistics, clinical trials statistics and medical research facilitation. Although the sample size was limited, the evaluation was carried out mostly at the qualitative level using interviews and performance observations, supplemented by questionnaires, ensuring that the results are still valuable.[8] The major aim of this usability study was to evaluate the suitability of the TrialViz searching facilities and visualization tool for use in a variety of research contexts in addition to its primary purpose in RCT. The participants had varying levels of previous experience working with either CPRD or The Health Improvement Network (THIN) primary care database and were recruited from the CPRD itself, University of Sussex, University College London and King's College London. None of the participants had been involved in the user group or had seen the system before.

The evaluation was based on the human computer interaction design principles such as visibility, feedback, constraints, affordance and Nielsen's heuristic evaluation guidance.[9 10] Users were asked to perform a set of typical research scenarios, during which they were observed, followed by interview and questionnaires commenting on their experience. Most of the output was qualitative; however, from the quantitative perspective we have scored users attempting scenarios on a Likert-type psychometric scale: 1=I cannot perform this scenario at all' to 5='I can perform this scenario easily without any help'. The user interface and visualization tool were also tested using the questionnaire on levels of satisfaction, learnability and functionalities as follows: Level of satisfaction: 1=very dissatisfied, 2=somewhat dissatisfied, 3=neutral, 4=somewhat satisfied, 5=very satisfied.

Learnability: 1=hard, 2=somewhat hard, 3=neutral, 4=somewhat easy, 5=easy.

Functionality: 1=did not fit the description, 2=missing a number of important functions, 3=borderline usable, 4=mostly usable 5=fit the description fully.

The technology acceptance model questionnaire[11] was used to measure perceived usefulness (PU), perceived ease of use (PEOU) and behavioral intention (BI) of the participants to use the tool in the future.

RESULTS

Data abstraction

We developed a web portal that allows users to access the entire database and run complex queries interactively (figure 2) and to visualize the results. This interface is based on 'stacks and cards'. Each card represents the results of a single query and the stacks represent a container of cards in which users can place multiple cards. The stack represents a logical combination of these cards and has two purposes: it is a visual presentation of 'set theory' rules and a means to implement the selection of date ranges. A search is built up using these cards. Cards are essentially patient lists representing longitudinal clinical events. They can be based on code lists, which the user can create themselves, or upload and, if required, modify existing 'public' code lists provided by other users or the CPRD (this library of code lists is an important component of the system). There is a search facility that allows users, building or modifying code lists to search for terms or Read codes (eg, 'diab' for diabetes or Read codes beginning with 'C'). Certain cards can also be based on test criteria and clinical history (eg, HbA1c or BMI) or indeed on the results of previous searches (as each search will result in a patient list which can be represented as a card).

Cards are logically grouped on stacks and (depending on the type of search) stack represents a union (OR) of the cards it contains while the search represents the intersect (AND) of the participating stacks. Conversely, a second search type intersects the cards on a stack and UNIONs the stacks in the search. There is also the ability to take the logical NOT of a stack and the user can therefore logically end up with a set of patients who did NOT have a particular medical event on a particular date. Users can create many stacks on the screen, into which they can place multiple cards. They can specify various demographic filters (eg, the age range of the patient) and the date range for events. All stacks have both absolute and relational date matching. This allows the cards on the stack to have occurred relative to another stack as well as occurring absolutely in time.

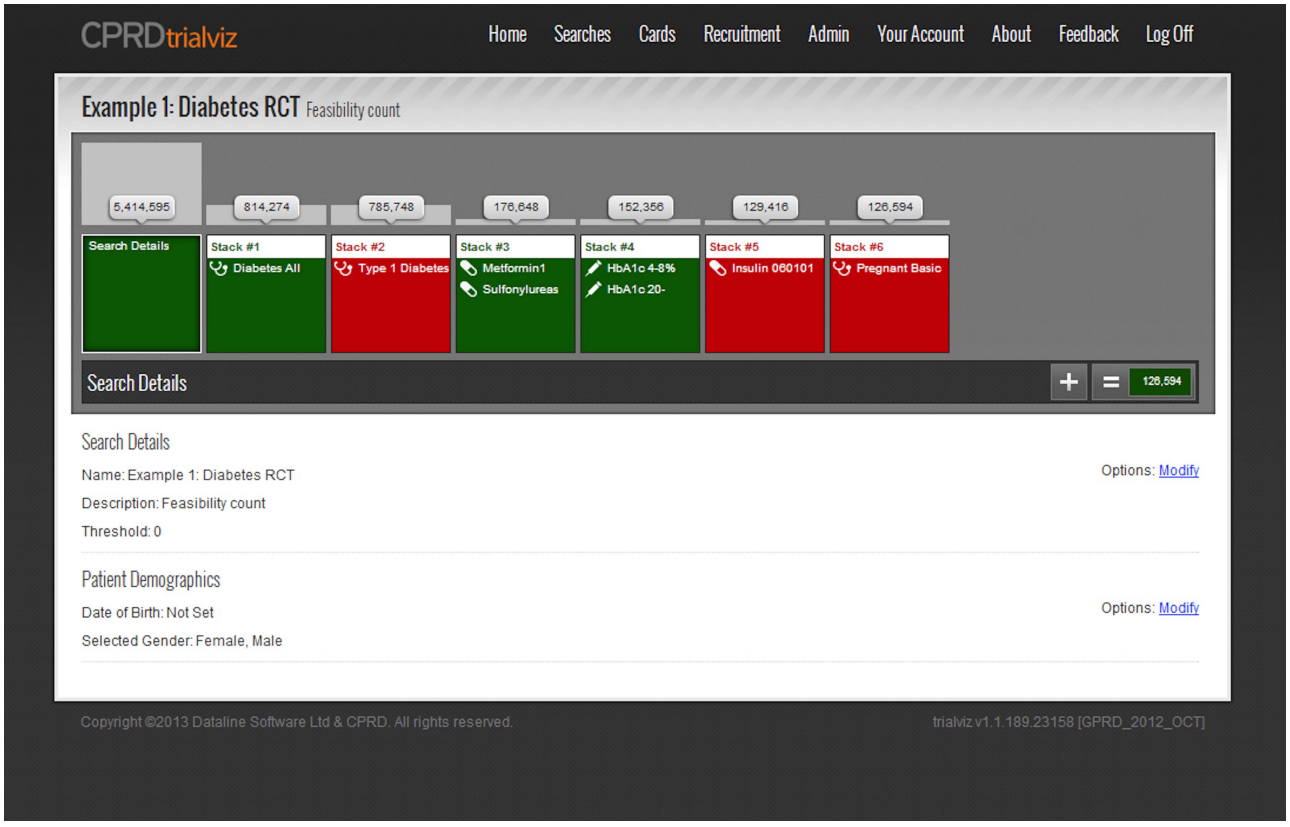

Figure 2 The stack and card interface showing the results of a search for the diabetes example. The counts are cumulative and represent how many patients are left within the search. RCT, randomized controlled trial.

Figures 2 and 3 show the results of running a query to select diabetes patients who do NOT have type 1 diabetes, AND who have been prescribed either metformin OR sulfonylureas (within the specified time), AND who have a specified level of HbA1c, AND who have NOT been prescribed insulin and who are NOT pregnant. Green stacks represent inclusion and red stacks exclusion criteria. The counts for each cumulative result are shown above the stacks, and the final patient count is shown by region in the map (figure 3). The colors give a rough indication of the average number of eligible patients per practice in that region, with blue representing smaller numbers, and red larger numbers.

The counts represent actual patient lists within the dataset, representing living patients who adhere to the search criteria at the time of the latest practice data extraction contributing to the CPRD version in use (updated monthly). As such, their CPRD identifiers represent potential individuals available for RCT recruitment selection—given that they are: (a) registered at practices that contribute data to CPRD; (b) have not exercised their right to opt out of data collection; and (c) can be contacted by the GP as an initial step in recruitment. CPRD data are collected and used in accordance with fully approved governance rules.

The CPRD identifiers are a pseudonymized version of patient system identifiers; we can initiate a recruitment process by sending these to participating practices. This recruitment is possible now and is functioning successfully in two CPRD-led proof of concept pragmatic RCTs.[12] This methodology has been subject to full ethical review and within the context of these studies is good clinical practice compliant. We are confident therefore that recruitment based on output from TrialViz will be acceptable for use in RCT due to the conceptual similarity of the patient identification process with these approved studies.

Data extraction

The TrialViz system is based on a three-database model (figure 4): (1) The TrialViz user database, which holds user searches, stack and card pointers and also contains the functional interface (query engine) for the TrialViz application. (2) The TrialViz cached results database, which contains volatile result sets for all cards, stacks and searches. The query engine checks this cache to see if a result sets exists (if it does, result sets are returned without continuing) before querying the CPRD SQL database and building new cached result sets. (3) The CPRD SQL database, which is a read-only SQL repository containing all the CPRD anonymized coded patient event data; this database is queried via user database and returns result sets to be cached.

We have leveraged the fast input/output speed of solid state drives to create an optimized SQL-based query-processing tool. The optimization is based on a divide and execute approach, and in conjunction with the database model and the underlying architecture it reduces the search time to seconds rather than hours or days. Patent is pending (applied for by Dataline Software) on innovative search algorithms, which allowed this impressive improvement in query execution time. Search time is also reduced by using cached results, so if the result set is already in the cache, the search time will be virtually instant.

We carried out a test of the performance of our query-processing algorithm versus existing CPRD performance for a typical simple query: create an intersection of three stacks, containing: (1) patients with asthma-related codelist in clinical events (returns 10 399 427 rows); (2) patients with arthritis-related codelist in clinical events (618 728 rows) and (3) patients with aspirin-related codelist in therapy events (32 516 196 rows). There was a dramatic increase in the performance when

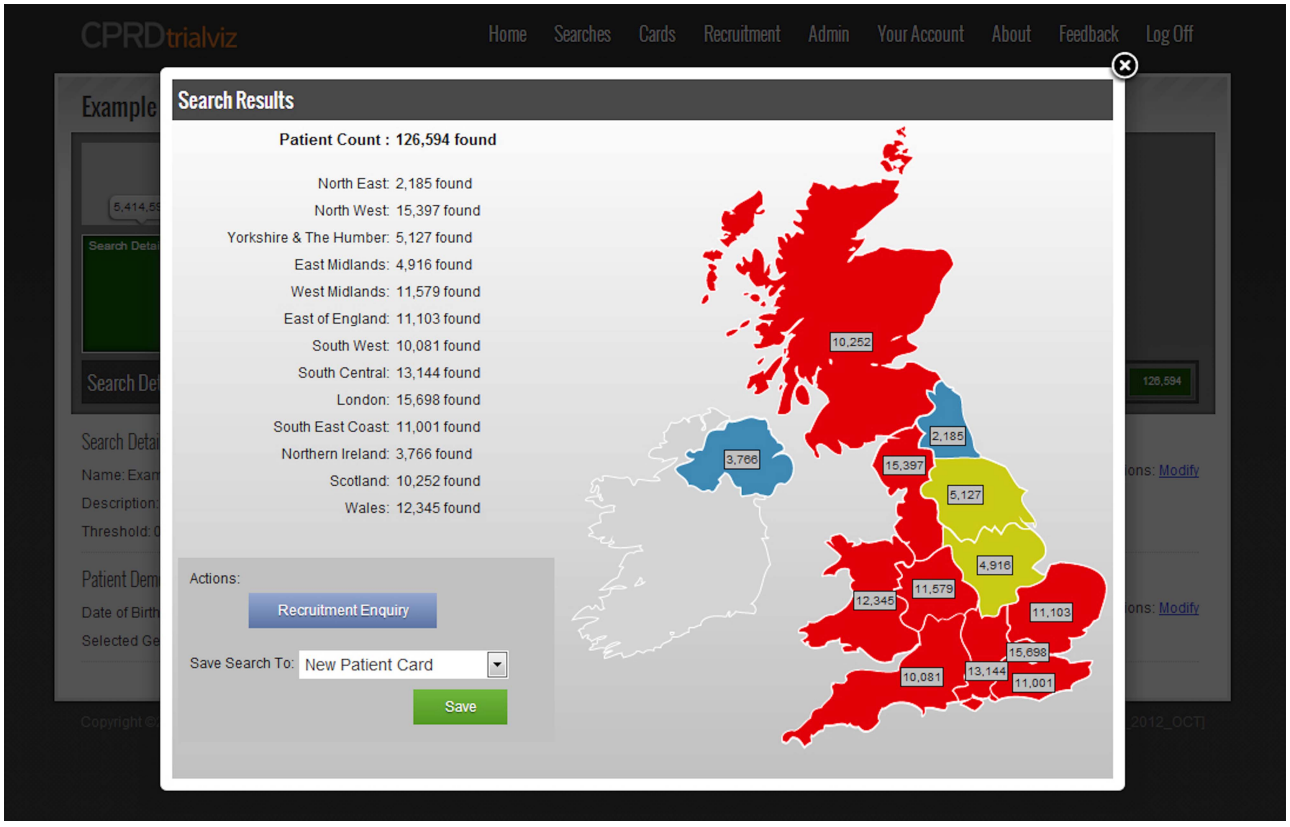

Figure 3 Visualization of patient number on map for the diabetes example.

our new TrialViz query-processing tool was used: processing time is reduced from nearly 5 h to 2 s.

Initial investigations indicate that TrialViz, due to its speedy, near real-time response, compares favorably to other similar (but much smaller scale) systems with text-based interfaces, such as IMS disease analyser,[13] ePCRN[14] and FARSITE;[15] however, a full parameterized benchmarking study was not possible within the framework of this project.

Data quality assurance

The recording of most measures improved significantly over time, with a noticeable improvement in some measures in 2004 when the UK NHS quality outcomes framework was introduced. In general, data quality in recent years appears to be good.[16] However, there were large variations between practices and nearly all practice-based variables had very skewed distributions, with several outliers. This is demonstrated by an example (figure 5), which shows the distributions of the percentage (for each practice) of patients with diabetes who have a code for type of diabetes. Inter-correlations between variables representing different categories of measures were generally very weak, and practices that were poor at recording for one aspect of quality were almost always very good at recording for other aspects. Due to the lack of correlations between the practice variables and inconsistency of poor performance across measures we decided not to use sophisticated multivariate methods for combining data quality measures into scores (as was first planned) but to take a more pragmatic approach and, after carrying out basic database checks for key measures common to most studies, to tailor the data quality indicators to the disease areas of interest. We are currently developing generic programs, which will enable users to generate sets of data quality indicators for each contributing GP practice to indicate their level of suitability for a particular research study. We plan to incorporate these indicators into TrialViz in due course. In addition, at the request of our users, we have compiled metadata, which include the unit used, the medians and range, for all the test results, plus boxplots showing the relative distributions for each unit used, which we shall make available as an online manual to users of the system.

User evaluation

Users were asked to perform the following typical scenarios:

1. Create a new search that fits the following criteria: (a) How many women aged 25–45 years are in the database, who have been diagnosed with ovarian cancer; (b) how many women have been diagnosed with ovarian cancer and had abdominal pain.
2. Create a new search answering how many patients are born before the year 2000 and had an event of either abdominal pain or abdominal distension.
3. Create a new 'Medical codelist card' that contains the Read code C10.00 (diabetes mellitus) and find out how many people aged 18 years and older have been diagnosed with it.

The average score for the scenarios performed on TrialViz Beta V.3.1 was 4.89 = 'I can perform this scenario easily without any help/ minimal help'.

We also undertook a technology acceptance model analysis based on users' interview results:[11]

$$\mathrm{PA} = \mathrm{PU} + \mathrm{PEOU}$$
$$\mathrm{BI} = \mathrm{PA} + \mathrm{PU},$$

where the coefficients PU and PEOU were derived from the

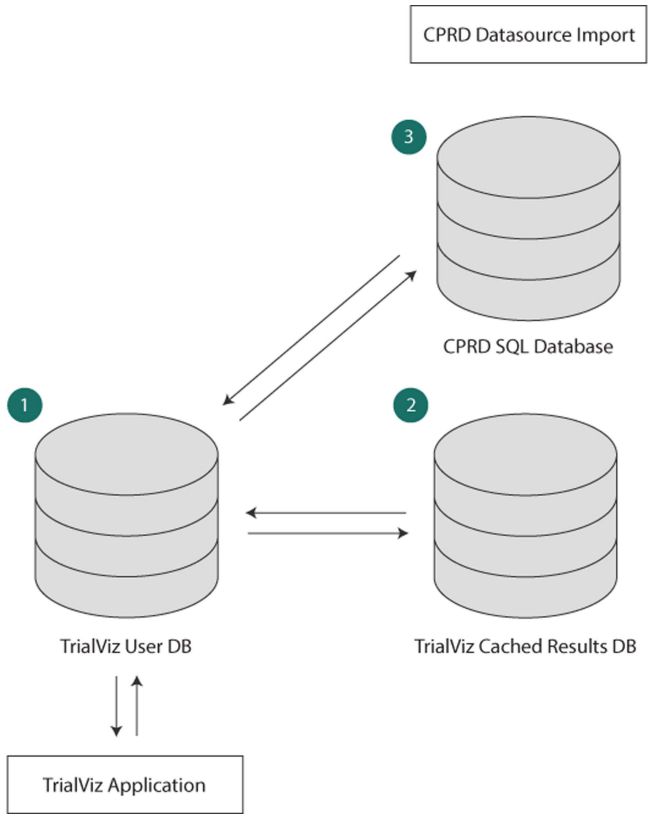

Figure 4 The three-database model that was used for TrialViz. CPRD, Clinical Practice Research Datalink; DB, database.

evaluation questionnaire (see table 1) and cross-referenced with interviews. As we can see from table 2, personal attitude (PA) towards the tool was very positive (derived from Likert-based scores from PU and PEOU), influencing the possibility to adopt the tool in the future in their work activities BI.

As shown in tables 1 and 2, users were satisfied overall with TrialViz and indicated a strong interest in adopting TrialViz in their work in the future. Open-ended questions also indicated that the users found TrialViz's search interface and code list facilities intuitive, memorable and easy to learn. They were also impressed with the range of functionality offered. The most praised feature of the tool was its fast performance of searches, unmatched by other currently available medical database search tools. However, the users indicated that visualization of search results could be more extensive and that a feature of exporting search results to a variety of formats would be useful.

DISCUSSION

Protocols for RCT are typically based on a large number of inclusion and exclusion criteria and the queries may be very complex. The flexible design, and three-database model (only one of which is dependent on the core database) means that it can be easily adapted for many other applications. By working on a generic core dataset, TrialViz could be developed to work with any large-scale electronic data. Such data resources are now a major focus within healthcare research and with their increased use the importance of the core data quality methodology will rise. Consequently, this will open up a plethora of parallel applications for the generic approach being developed within this project. Apart from the more obvious applications for health services research, current additional applications include the use of data quality scores in terms of observational data studies to model uncertainty in data. The application of the methodology per se is important in obtaining the strategies to assess the quality and usefulness of routinely collected data in terms of various spheres of research. Furthermore, sister systems to the one proposed could be easily developed to facilitate feedback to the data collecting entities (general practices in this

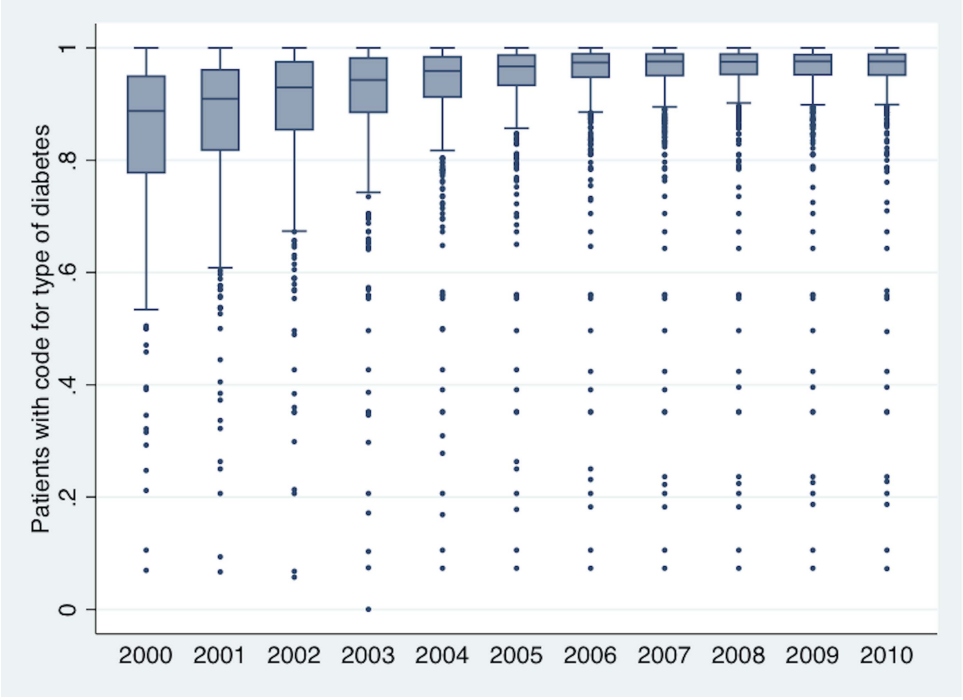

Figure 5 Boxplots showing the distribution of the proportion of patients (per GP practice) with a first diagnosis code for diabetes in each year whose type had been recorded at some point in their record.

Table 1 Questionnaire results for general usability of the TrialViz tool

System	Level of satisfaction (1=least–5=very)	Learnability (1=hard–5=easy)	Functionalities (1=did not fit description–5=fit description)
TrialViz	3.91	4.38	4.58

Table 2 TAM analysis for TrialViz tool.

System	PU (maximum 5)	PEOU (maximum 5)	PA (maximum 10)	BI (maximum 15)
TrialViz	3.67	4.5	8.17	11.84

BI, behavioral intention; PA, personal attitude; PEOU, perceived ease of use; PU, perceived usefulness.

case), which may act as a positive feedback mechanism for the enhancement of data quality at source.

CONCLUSIONS

In this paper we presented the development of a system (TrialViz), which allows users to interrogate a large database of patient records and to visualize the results of complex queries in near real time. This represents a considerable advance over most other currently available systems. Our query-processing tool offers opportunities for exploring the data in ways that could never have been envisaged before and also delivering information in real or near real time. The TrialViz system identifies individuals potentially available for study recruitment, representing a significant advantage over other systems that investigate general rates of disease or patient availability within geographical areas. TrialViz facilitates interaction with potential study participants via their GP. TrialViz was developed independently from the development of CPRD as an organization; however, its scope aligns well with CPRD objectives and as such will be incorporated into the suite of tools taken forward under the umbrella of CPRD services.

Contributors ART, TVS, TW, NB, AB and JW made substantial contribution to conception and design; TVS, TW, SP and BAR to acquisition of data and TVS, TW, SP, BAR and ART to analysis and interpretation. ART, AB, JW and NB drafted parts of the article, and all authors revised it critically for important intellectual content.

Funding This work was supported by the Technology Strategy Board 'Harnessing large and diverse sources of data' grant number 100926. ART currently holds a fellowship funded by the CPRD.

Competing interests TW, SP and TVS are employees within the CPRD group of the Medicines and Healthcare Products Regulatory Agency (MHRA). ART is funded by CPRD. NB, BAR, JW and AB have no competing interests.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Williams T, VanStaa T, Padmanabhan S, *et al*. Recent advances in utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Ther Adv Drug Saf* 2012;3:88–99.
2 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *JAMIA* 2013;20:117–21.
3 PRIMIS. Information Guides for Practices. Read Codes Version 2 (5-byte); 2007. http://www.nottm-vts.org.uk/ReferenceMaterial/
4 Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiol Drug Saf* 2009;18:704–7.
5 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010;60:199–206.
6 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144–51.
7 Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM* 2002;45:211–18.
8 Sue VM, Ritter LA. *Conducting online surveys*. Sage Publications, 2007.
9 Rogers Y, Sharp H, Preece J. *Interaction design: beyond human—computer interaction*. Wiley, 2011.
10 Nielsen J. *Usability engineering. Morgan Kaufmann series in interactive technologies*. Morgan Kaufmann, 1994.
11 Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. *Manage Sci* 1989;35:982–1003.
12 van Staa TP, Goldacre B, Gulliford M, *et al*. Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* 2012:344.
13 Bate A, Noren NG, Star K, *et al*. Results from data mining in IMS disease analyzer patient records [Meeting Abstract]. *Pharmacoepidemiol Drug Saf* 2007;16:S256.
14 Peterson K, Fontaine P, Speedie S. The electronic Primary Care Research Network (ePCRN): a new era in practice-based research. *J Am Board Fam Med* 2006;19:93–7.
15 Ainsworth J, Buchan I. Preserving consent-for-consent with feasibility-assessment and recruitment in clinical studies: FARSITE architecture. In: Solomonides T, HofmannApitius M, Freudigmann M, Semler SC, Legre Y, Kratz M. eds. *Healthgrid research, innovation and business case. vol. 147 of studies in health technology and Informatics*. IOS Press, 2009:137–48.
16 Tate AR, Beloff N, Puri S, *et al*. Developing quality scores for electronic health records for clinical research: a study using the General Practice Research Database. In: *ACM Proceedings of MIXHS11*; 28 October 2011, Glasgow, Scotland.

Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface

A Rosemary Tate, Natalia Beloff, Balques Al-Radwan, et al.

J Am Med Inform Assoc published online November 22, 2013
doi: 10.1136/amiajnl-2013-001847

Notes